Comparison of network inference packages and methods for multiple network inference

Nathalie Villa-Vialaneix[a], **Nicolas A. Edwards**[b], **Laurence Liaubet**[b]
Nathalie Viguerie[c], **Magali SanCristobal**[b]

[a]Laboratoire SAMM - Université Paris 1 (Panthéon-Sorbonne)
90 rue de Tolbiac, 75013 Paris - France
nathalie.villa@univ-paris1.fr

[b]INRA, UMR444 - Laboratoire de Génétique Cellulaire
F-31326 Castanet Tolosan cedex, France
nicolas.ae@free.fr, {laurence.liaubet,magali.san-cristobal}@toulouse.inra.fr

[c]Inserm UMR1048, Obesity Research Laboratory
I2MC, Institute of Metabolic and Cardiovascular Diseases
CHU Rangueil, Toulouse
nathalie.viguerie@inserm.fr

Keywords: network inference, transcriptomic data, gene co-expression network, Gaussian graphical model, multiple graphical structure

Integrative and systems biology is a very promising tool for deciphering the biological and genetic mechanisms underlying complex traits. In particular, gene networks are used to model interactions between genes of interest. They can be defined in various ways, but a standard approach is to infer a co-expression network from genes expression measured by means of sequencing techniques (for example, microarrays). Among methods used to perform the inference, **Gaussian graphical models** (GGM) are based on the assumption that the gene expressions are distributed as Gaussian variables, and Σ is their covariance matrix. Non-zero partial correlations between two genes are modeled by network edges, and are directly obtained from the inverse of Σ. But it turns out that estimating the inverse of Σ leads to an ill posed problem, since this kind of data leads to a number of observations (typically less than one hundred) that is usually much smaller than the number of variables (the number of genes/nodes in the network can range from a few hundred to several thousands). To overcome this difficulty, the seminal papers [8, 9] were the basis for the R package **GeneNet**, in which the partial correlation is estimated either by means of a bootstrap approach (not available in the package anymore) or of a shrinkage approach. More recently, the ability to handle genomic longitudinal data was also added as described in [7]. Then, [6] and later [3] introduced sparse approaches, both implemented in the R package **glasso** (graphical LASSO). Similarly, [4] describes the methods implemented in the package **parcor** that provides several regularization frameworks (PLS, ridge, LASSO...) to infer networks by means of Gaussian graphical models. Finally, [2, 1] describe several extensions of the Gaussian graphical model implemented in the package **simone** such as latent variable models and time-course transcriptomic data.

In systems biology, an interesting issue is to link gene functioning to an external factor. Thus, transcriptomic data are often collected in different experimental conditions. One must then understand which genes are correlated *independently* from the condition and which ones are correlated *depending* on the condition, under the plausible biological assumption that a common functioning should exist regardless of said condition. A simple naive approach would be to infer a different network from each sample, and then to compare them. Alternative approaches

are described in [2, 1] and implemented in **simone**: the log-likelihood can be penalized by a modified group-LASSO penalty or the empirical covariance matrix can be modified by adding a component depending on all samples. The purpose of this communication is to present a full comparative case study of this problem on two real data sets.
The first dataset has been collected during the DiOGenes project[1]: a few hundreds human obese individuals were submitted to a 8 weeks low calorie diet. The expressions of pre-selected genes as well as physiological variables (age, weight, waist size...) were collected *before* and *after* the diet (see [5] for further information). The underlying issue is to understand how the diet has affected the correlations between all these variables. The second data set has been collected during the Delisus project[2]: the expression of several thousands genes were collected from 84 pigs (in both *Landrace* and *Large White* breeds). The underlying issue is to understand how the breed affects the correlations between a set of selected genes which were found to be differentially expressed for the breed.
The comparison is lead by using independent inference from the packages **GeneNet**, **glasso** and **simone** or by using the different joint models included in **simone** or even by proposing new joint approaches based on the aforementionned packages. Networks are inferred from the previously described real datasets or from simulated datasets that mimic the real ones. The proximity between networks inferred from different methods or from different conditions is assessed by means of common edge counts, or, when available, by the accuracy of the inferred network when compared to the true one. A biological discussion about the relevance of the inferred networks will also be provided.

References

[1] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

[2] J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25(3):417–418, 2009.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[4] N. Kraemer, J. Schaefer, and A.L. Boulesteix. Regularized estimation of large-scale gene regulatory networks using Gaussian Graphical models. *BMC Bioinformatics*, 10:384, 2009.

[5] T.M. Larsen, S.M. Dalskov, M. van Baak, S.A. Jebb, A. Papadaki, A.F.H. Pfeiffer, J.A. Martinez, T. Handjieva-Darlenska, M. Kunešovà, M. Pihlsgård, S. Stender, C. Holst, W.H.M. Saris, and A. Astrup. Diets with high or low protein content and glycemic index for weight-loss maintenance. *New England Journal of Medicine*, 363:2102–2113, 2010.

[6] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistic*, 34(3):1436–1462, 2006.

[7] R. Opgen-Rhein and K. Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65, 2006.

[8] J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[9] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:1–32, 2005.

[1] supported by funding from the European Communities (DiOGenes, FP6-513946, MolPAGE, LSHG-CT-2004-512066 and ADAPT, HEALTH-F2-2008-2011 00), Fondation pour la Recherche Médicale and Région Midi-Pyrénées `http://www.diogenes-eu.org`

[2] funded by the ANR, `http://www.inra.fr/les_partenariats/programmes_anr/genomique/genanimal/appel_a_projets_2007/delisus`